УДК 621.391

Ткаченко М. В., к.т.н.[1], (ORCID: 0000-0003-2929-3495);
Федоренко Р. М., к.е.н.[1], (ORCID: 0000-0003-2929-3495);
Кондратенко Ю. В.[2], (ORCID: 0000-0003-2929-3495);
Зотова І. Г.[2], (ORCID: 0000-0002-8804-3085)

[1] – Київський національний університет імені Тараса Шевченка, Київ;
[2] – Центр воєнно-стратегічних досліджень Національного університету оборони України імені Івана Черняховського, Київ

Методи автоматичної ідентифікації диктора за голосом

Резюме. Проведено аналіз методів автоматичного розпізнавання диктору за голосом, на підставі якого здійснено вибір методу для рішення задачі текстонезалежного розпізнавання.

Ключові слова: мовний сигнал, диктор, розпізнавання, динамічна трансформація часової шкали, приховані марковські процеси, векторне квантування, опорні вектори, гаусові суміші.

Постановка проблеми. Кожна людина має індивідуальні голосові характеристики, які визначаються особливостями будови його голосових органів [1]. У процесі спілкування люди здатні на підсвідомому рівні розрізняти голоси інших людей, однак для обчислювальної техніки ця задача є нетривіальною і вимагає цілеспрямованих досліджень.

Аналіз останніх досліджень і публікацій. Задачу розпізнавання особистості за голосом було поставлено більш 40 років тому, але дослідження в цій області тривають і досі. За останні роки спостерігається значне підвищення якості розпізнавання мовної інформації, однак основна проблема автоматичного розпізнавання диктора в будь-якому середовищі все ще далека від вирішення [1-7].

Мета статті - аналіз існуючих методів розпізнавання мовної інформації, визначення їх слабких і сильних сторін для обгрунтування вибору найбільш сприйнятливого стосовно розпізнавання диктора за голосом.

Виклад основного матеріалу. Перший пристрій для розпізнавання мови з'явився ще в 1952 році, він міг розпізнавати вимовлені людиною цифри. [1] У 1962 році на ярмарку комп'ютерних технологій в Нью-Йорку було представлено пристрій IBM Shoebox, яки був здатний розпізнавати 16 вимовлених слів і цифри від 0 до 9.

Комерційні програми з розпізнавання мови з'явилися на початку дев'яностих років. Зазвичай їх використовують люди, які через травми кінцівок не в змозі набирати велику кількість тексту. Ці програми (наприклад, Dragon Naturally Speaking, Voice Navigator) трансформують голос користувача в текст, таким чином, розвантажуючи його руки. Надійність перекладу у таких програм не дуже висока, але з роками вона поступово покращується.

Світовий ринок розпізнавання мови є одним з найбільш швидкозростаючих ринків в голосовий індустрії. Велика частина зростання на ринку обумовлена продукцією, що приходить з Америки, Європи, Близького Сходу, Африки (EMEA) і Азіатсько-Тихоокеанського регіону (ATP), яка призначена для в використання у галузях охорони здоров'я і надання фінансових послуг, а також у державному секторі.

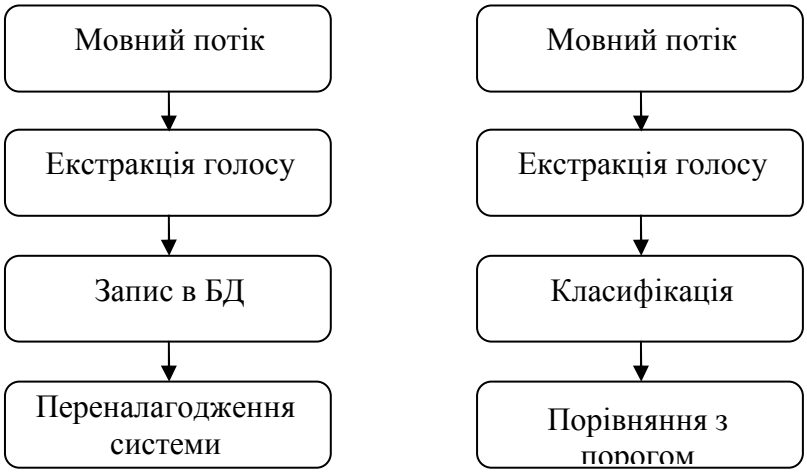
Зростання світового ринку розпізнавання голосу залежить від множини факторів. Одним з основних факторів є збільшення попиту на послуги голосової біометрії. Зі збільшенням складності і частоти порушень безпеки, остання продовжує залишатися одним з основних вимог для Збройних Сил України. Високий попит голосової біометрії, яка є унікальною для будь-якої людини, має вирішальне значення у встановленні особи людини.

Військові відомства в більшості країн використовують вкрай обмежені зони для того, щоб запобігти проникненню зловмисників. Для забезпечення секретності і безпеки в цій зоні, військові використовують системи розпізнавання голосу.

Ці системи допомагають військовим установам виявляти наявність несанкціонованих проникнень в захищену зону. Система містить базу даних голосів військовослужбовців і державних чиновників, які мають допуск до захищеної території. Ці люди ідентифікуються системою розпізнавання голосу, тим самим запобігається допуск людей,

чиїх голосів немає в базі даних системи. На додаток можна сказати, що ВПС США використовують голосові команди для керування літаком. Крім того, військові відомства використовують розпізнавання мови і систему Voice-to-text для комунікації з громадянами в інших країнах. Наприклад, американські військові активно використовують системи розпізнавання мови в операціях в Іраку і Афганістані. Таким чином, існує високий попит на розпізнавання мови і голосу для військових цілей.

Будь-яка система розпізнавання працює в двох режимах: в режимі *реєстрації* та режимі *ідентифікації*. Іншими словами, необхідно мати приклад голосу.



а) реєстрація **б)** ідентифікація

Рис. 1. Типова схема роботи системи в режимах *реєстрації* та

На рис. 1 представлена загальна схема роботи системи в кожному з режимів. Як можна побачити, ці режими досить схожі. Обом для роботи необхідно захопити мовний аудіо потік і обчислити його основні ознаки. Відмінність полягає у порядку реагування на ці ознаки. При реєстрації їх, так чи інакше, необхідно запам'ятати для використання в майбутньому, адже набагато ефективніше працювати з основними ознаками, ніж з вихідними необробленими даними. При ідентифікації нічого зберігати не можна, так як система на даному етапі не має зворотного зв'язку і не може достовірно знати приналежність голосу.

Ознаки являють собою набір чисел, що характеризують диктора - вектор в багатовимірному просторі. А завдання класифікатора полягає в тому, щоб побудувати функцію відображення цього багатовимірного простору в простір дійсних чисел. Іншими словами, його завдання полягає в отриманні числа, яке б характеризувало міру схожості.

$D = \left(\sum_{i=1} |x_i| \right)^{1/2}$ - евклідова відстань (Euclid);

$y = \langle w, x \rangle + w_0$ - розділяюча гіперплощина (SVM):

$U_i = f\left(\sum_{i=1}^{n} w_{ji} x_i + w_{j0} \right)$ - мережа нелінійних функцій (MLP)

Існують наступні проблеми і обмеження задачі розпізнавання особистості за голосом, які слід враховувати при побудові рішення:

емоційний стан диктора;

складна акустична обстановка (шуми і перешкоди);

різні канали зв'язку при навчанні та розпізнаванні нейронних мереж;

природні зміни голосу диктора.

Мовлення представляє собою складний сигнал, що утворюється в результаті перетворень, які відбуваються на кількох рівнях: семантичному, мовному, артикуляційному (рівні голосового апарату людини) і акустичному (рівні фізичних властивостей звуку). Відмінності в цих перетвореннях тягнуть за собою відмінності у властивостях мовного сигналу. При вирішенні задачі розпізнавання диктора за голосом всі ці відмінності можуть бути використані для того, щоб виділити індивідуальні характеристики голосу кожної людини.

Незважаючи на те, що методи розпізнавання багато в чому відрізняються, в цілому можна виділити наступні основні етапи оброблення даних , характерні для кожного з відомих методів:

- витяг ознак з вхідного мовного сигналу;

- побудова моделі (шаблону) диктора на основі отриманих на попередньому кроці векторів ознак.

Dynamic Time Warping (DTW) - метод динамічної трансформації часової шкали дає змогу знайти близькість між двома послідовностями вимірювань за деякий проміжок часу. Вперше цей метод був застосований в розпізнаванні мови для визначення того, як два мовних сигнали представляють одну і ту ж вихідну виголошену фразу. У загальному випадку ці послідовності можуть бути різної довжини, і вимірювання можуть проводитися з різною швидкістю [2].

Алгоритм DTW обчислює матрицю розмірності $(M \times N)$, де *M* - кількість векторів - характеристик образу сигналу, що розпізнається, і *N* - кількість векторів еталону. Матричний елемент *D (i, j)* - є оцінкою глобального шляху вирівнювання до точки *(i, j)*. Процес обчислення описується формулами:

$$D(i,j) = \min[D(i-1,j), D(i-1,j-1), D(i,j-1)] + d(i,j),$$

$D(i,j) = d(i,j)$, де $d(i,j)$ – це локальна оцінка в точці (i,j).

Основною перевагою алгоритму DTW є простота реалізації. Проте, даний алгоритм непридатний для вирішення задачі текстонезалежної ідентифікації диктора.

Hidden Markov Model (HMM) – прихована марковська модель - статистична модель, яка може використовуватися для вирішення задачі класифікації прихованих параметрів на основі спостережуваних даних. HMM – це кінцевий автомат, в якому переходи між станами здійснюються з певною ймовірністю, і задано стартовий стан, з якого починається процес. Через дискретні моменти часу може здійснюватися перехід в нові стани. При цьому кожному прихованому стану з заданою ймовірністю відповідає стан, що спостерігається. Крім того, поточний стан автомата залежить тільки від кінцевого числа попередніх, а закон зміни станів не змінюється в часі [3].

HMM мають досить високу точність розпізнавання, але, як і DTW, застосовуються в основному для задач текстозалежної ідентифікації диктора.

Vector Quantization (векторне квантування) – розбиття простору можливих значень векторної величини на кінцеве число областей. Цей метод обробки сигналу дає змогу моделювати щільність ймовірності, функції розподілу векторів. Спочатку цей метод використовувався для стиснення даних. Він працює шляхом поділу великого набору векторів на групи, що мають приблизно однакові значення. У методі векторного квантування вибірка з навчальних векторів перетворюється у фіксовану множину кодових векторів. Одним з поширених методів формування подібної множини, званого також кодовою книгою, є алгоритм *K*-середніх.

Алгоритм *K*-середніх розбиває вихідну множину на *K* кластерів, де *K* - попередньо задане число. Для цього, спочатку значення середніх ініціюються деякими векторами з вихідної множини. Потім на кожній ітерації алгоритму відбувається розподіл векторів в найближчі до них кластери (для цього обчислюється відстань між вектором і поточними значеннями середніх) і перерахунок середнього в кожному кластері. Алгоритм завершується після того, як на черговій ітерації стани кластерів не змінилися або після досягнення заданої максимальної кількості ітерацій.

Метод векторного квантування простий в реалізації та може бути застосований для вирішення задачі текстонезалежної ідентифікації диктора, проте не завжди дає високу точність розпізнавання.

Support vector machine (метод опорних векторів) – метод полягає в побудові оптимальної поділяючої гіперплощини. Під оптимальною розуміється гіперплощина, яка перпендикулярна найкоротшому відрізку, що з'єднує опуклі оболонки різних класів, і проходить через середину цього відрізка. Іншими словами, оптимальна гіперплощина повинна максимізувати ширину поділяючої смуги між класами.

Метод на опорних векторах являє собою особливий вид класифікатора, який знаходить відповідність між вектором даних x_l та одним з двох класів $y_l\{-1,+1\}$ у відповідності зі знаком виразу:

$$y = \sum_{l=1}^{L} \lambda_l y_l K(x_l, x) - b,$$

де $K(x_l, x)$ – симетричне позитивно-певне ядро інтегрального рівняння, яке підпорядковується ряду обмежень [5], оскільки є скалярним добутком в певному просторі:

$$K(x_l, x) = \phi(x_l)^T \cdot \phi(x),$$

де ϕ - функція, яка переводить вектор вихідних даних в простір з більшою (можливо нескінченної) розмірністю.

В якості ядра часто виступають поліноміальні функції, радіальна базисна функція і сігмаідальна функція. Параметри λ_i і b визначаються виходячи з рішення задачі квадратичного програмування з лінійними обмеженнями [6]. Основна перевага методу ґрунтується на тому, що тільки мала частина коефіцієнтів λ_i відмінна від нуля. Іншими словами, тільки малий набір опорних векторів (x_i з ненульовими λ_i) потрібен для класифікації згідно з виразом (1).

Метод опорних векторів дає високу точність класифікації, має теоретичне обґрунтування, дає змогу застосовувати різні підходи до класифікації згідно з вибором функції. Серед недоліків слід відзначити проблему повільного навчання нейромережі в разі завдання багатокласового розпізнавання.

Gaussian Mixture Model (модель гаусових сумішей) представляє собою параметричну функцію щільності ймовірності. Дана модель є вдалою варіацією стохастичної моделі для побудови систем розпізнавання [4]

$$f\left(X^k, w\right) = \sum_{i=1}^{k} p_i G\left(w|x_i, \Gamma_i\right),$$

де $G\left(w|x_i, \Gamma_i\right)$ – щільність гаусівського розподілу із середнім x_i та коваріаційною матрицею Γ_i , $i \in \overline{1,k}$

$$\Gamma_i = \beta_i U_i D_i U_i^T, i \in \overline{1,k}$$

де β_i – константа об’єму коваріаційної матриці;

U_i - матриця собистих векторів, що визначає орієнтацію кластеру;

$D_i = diag\left(\lambda_{1,i},...,\lambda_{d,i}\right)$ - матриця особистих чисел, що визначає форму коваріаційної матриці, $\lambda_{d,i} \le \lambda_{d-1,i} \le ... \le \lambda_{1,i} = 1$.

Модель зручна для моделювання характеристик голосу диктора, каналу звукозапису, навколишнього середовища. Кожна з компонент моделі відображає деякі загальні, але індивідуальні для кожного диктора особливості голосу. Саме тому даний підхід можна успішно застосовувати для вирішення задачі ідентифікації диктора.

Для того щоб побудувати модель диктора, необхідно точно оцінити її параметри, які найбільш точно відповідають розподілу векторів ознак навчального висловлювання. Існує певний ряд методів для оцінки параметрів моделі. Одним з найбільш популярних і добре себе зарекомендованих є метод оцінки максимальної правдоподібності [7]. Мета оцінки за даним методом полягає у визначенні параметрів моделі, які максимально підвищать ймовірність правдоподібності моделі при заданих даних для навчання.

Модель являє собою ефективний алгоритм, який дає змогу проводити ідентифікацію з високою точністю розпізнавання [8]. Однак виникає ряд проблем, які пов’язані з вибором числа компонентів моделі та ініціалізацією її початкових параметрів.

Висновки. На даний час існує певна кількість методів, що дають змогу вирішувати завдання текстонезалежної ідентифікації диктора за голосом, причому кожен із наведених методів має свої переваги та недоліки. Проте, найбільш поширеним методом є Gaussian Mixture Model. Моделі гаусових сумішей добре себе зарекомендували в якості стохастичної моделі для побудови систем розпізнавання [9]. Вони зручні не тільки для моделювання характеристик голосу диктора, але і каналу звукозапису, навколишнього середовища. Окремі компоненти моделі можуть моделювати окрему множину акустичних ознак. Кожна з компонент моделі відображає як загальні, так і індивідуальні для кожного диктора особливості голосу. Саме тому даний підхід можна успішно застосовувати для вирішення завдання текстонезалежної ідентифікації диктора.

Напрямки подальших досліджень. Ефективна система розпізнавання мови має враховувати такі етапи обробки вхідного сигналу, як видалення шуму, сегментація, виділення вокалізованих ділянок, параметризація, розпізнавання, коригування за словником з оберненим зв’язком. Зрозуміло, що жоден метод не може вповні охопити усі етапи. Ефективна система має поєднувати в собі найкращі методи виконання кожного етапу, використовуючи їх переваги. Тому в подальшому планується провести аналіз і відібрати ефективні методи обробки сигналу для створення відповідної системи.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Campbell J. P. Speaker Recognition: A Tutorial // Proceedings of the IEEE.1997. Vol. 85, № 9. P. 1437-1462.

2. Ing-Jr Ding, Chih-Ta Yen, Yen-Ming Hsu. Development so Machine Learning Schemes for Dynamic Time-Wrapping-Based Speech Recognition // Mathematical Problems in Engineering. 2013.
3. Daniel Ram age. Hidden Markov Models Fundamentals // CS229 Section Notes. 2007.
4. Методы автоматического распознавания речи: в 2-х кн. / под ред. У. Ли ; пер. с англ. О. В. Александровой ; под ред. А. А. Воронова. М.: МИР, 1983. – Кн. 1. – 328 с.
5. Boser B., Guyon I., Vapnik V. A training algorithm for optimal margin classifier // in Proc. Of the Fifth Annual ACM Workshop on Computational Learning Theory. 1992. P. 144–152.
6. Vapnik V.The Nature of Statistical Learning Theory / V. Vapnik, Springer, Second Edition, 1999.
7. Кульбак С. Теория информации и статистика. М.: Наука, 1967. 408 с.
8. X. Huang, A. Acero, H. Hon. Spoken language processing: a guide to theory, algorithm, and system development. – Prentice Hall PTR, 2001. P. 936.
9. Forum S. Digital Speech Processing, Synthesis and Recognition // Marcel Dekker, New York, 1989.

Стаття надійшла до редакційної колегії 22.11.2018

Ткаченко М. В., к.т.н.[1];
Федоренко Р. Н., к.е.н.[1];
Кондратенко Ю. В.[2];
Зотова И. Г.[2]

[1] – Киевский национальный университет имени Тараса Шевченко, Киев

[2] – Центр военно-стратегических исследований Национального университета оборони Украины имени Ивана Черняховского, Киев

Методы автоматической идентификации диктора по голосу

Резюме. Проведен анализ методов автоматического распознавания диктора по голосу, на основании котрого осуществлен выбор метода для решения задачи текстонезависимого распознавания.

Ключевые слова: речевой сигнал, диктор, распознавание, динамическая трансформация временной шкалы, скрытые марковские процессы, векторное квантование, опорные векторы, гауссовские смеси.

M. Tkachenko, PhD, (Technical)[1];
R. Fedorenko, PhD (Economic)[1];
Y. Kondratenko[2];
I. Zotova[2]

[1] – Kyiv National Taras Shevchenko University, Kyiv;

[2] – Center for Military and Strategic Studies of the National Defence University of Ukraine named after Ivan Cherniakhovskyi, Kyiv

Methods for automatic speaker identification by voice

Resume. The analysis of the methods of automatic recognition of the announcer by voice was carried out, on the basis of which a choice was made of the method for solving the problem of text-independent recognition.

Keywords: speech signal, speaker, recognition, dynamic transformation of the time scale, hidden Markov processes, vector quantization, reference vectors, Gaussian mixtures.